OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE

MATHEMATICS (MEI)

4767

Statistics 2

Candidates answer on the Answer Booklet

OCR Supplied Materials:
- 8 page Answer Booklet
- Graph paper
- MEI Examination Formulae and Tables (MF2)

Other Materials Required:
None

Monday 19 January 2009
Afternoon

Duration: 1 hour 30 minutes

INSTRUCTIONS TO CANDIDATES

- Write your name clearly in capital letters, your Centre Number and Candidate Number in the spaces provided on the Answer Booklet.
- Use black ink. Pencil may be used for graphs and diagrams only.
- Read each question carefully and make sure that you know what you have to do before starting your answer.
- Answer **all** the questions.
- Do **not** write in the bar codes.
- You are permitted to use a graphical calculator in this paper.
- Final answers should be given to a degree of accuracy appropriate to the context.

INFORMATION FOR CANDIDATES

- The number of marks is given in brackets **[ ]** at the end of each question or part question.
- You are advised that an answer may receive **no marks** unless you show sufficient detail of the working to indicate that a correct method is being used.
- The total number of marks for this paper is **72**.
- This document consists of **4** pages. Any blank pages are indicated.

1 A researcher is investigating whether there is a relationship between the population size of cities and the average walking speed of pedestrians in the city centres. Data for the population size, x thousands, and the average walking speed of pedestrians, $y\,\mathrm{m\,s^{-1}}$, of eight randomly selected cities are given in the table below.

x	18	43	52	94	98	206	784	1530
y	1.15	0.97	1.26	1.35	1.28	1.42	1.32	1.64

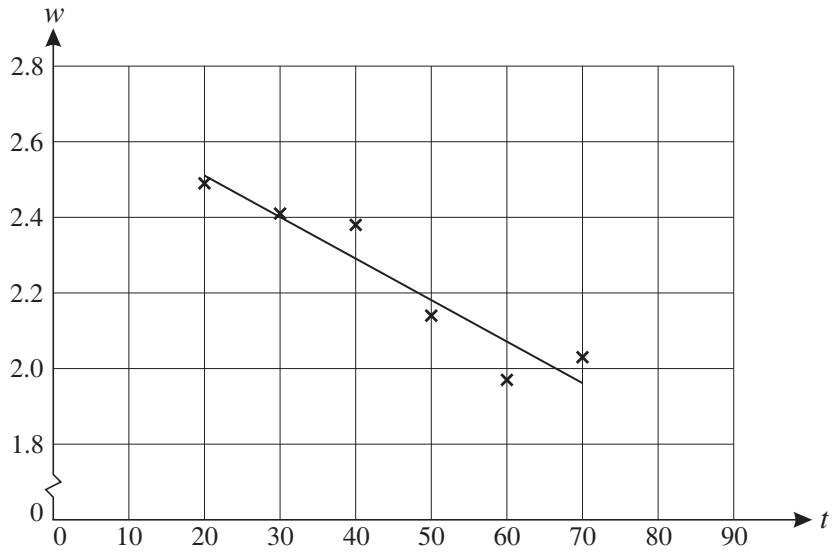
(i) Calculate the value of Spearman's rank correlation coefficient. **[5]**

(ii) Carry out a hypothesis test at the 5% significance level to determine whether there is any association between population size and average walking speed. **[6]**

In another investigation, the researcher selects a random sample of six adult males of particular ages and measures their maximum walking speeds. The data are shown in the table below, where t years is the age of the adult and $w\,\mathrm{m\,s^{-1}}$ is the maximum walking speed. Also shown are summary statistics and a scatter diagram on which the regression line of w on t is drawn.

t	20	30	40	50	60	70
w	2.49	2.41	2.38	2.14	1.97	2.03

$n = 6 \quad \Sigma t = 270 \quad \Sigma w = 13.42 \quad \Sigma t^2 = 13\,900 \quad \Sigma w^2 = 30.254 \quad \Sigma tw = 584.6$



(iii) Calculate the equation of the regression line of w on t. **[5]**

(iv) (*A*) Use this equation to calculate an estimate of maximum walking speed of an 80-year-old male. **[2]**

(*B*) Explain why it might not be appropriate to use the equation to calculate an estimate of maximum walking speed of a 10-year-old male. **[2]**

2 Clover stems usually have three leaves. Occasionally a clover stem has four leaves. This is considered by some to be lucky and is known as a four-leaf clover. On average 1 in 10 000 clover stems is a four-leaf clover. You may assume that four-leaf clovers occur randomly and independently.

A random sample of 5000 clover stems is selected.

(i) State the exact distribution of X, the number of four-leaf clovers in the sample. **[2]**

(ii) Explain why X may be approximated by a Poisson distribution. Write down the mean of this Poisson distribution. **[3]**

(iii) Use this Poisson distribution to find the probability that the sample contains at least one four-leaf clover. **[2]**

(iv) Find the probability that in 20 samples, each of 5000 clover stems, there are exactly 9 samples which contain at least one four-leaf clover. **[3]**

(v) Find the expected number of these 20 samples which contain at least one four-leaf clover. **[2]**

The table shows the numbers of four-leaf clovers in these 20 samples.

Number of four-leaf clovers	0	1	2	>2
Number of samples	11	7	2	0

(vi) Calculate the mean and variance of the data in the table. **[3]**

(vii) Briefly comment on whether your answers to parts **(v)** and **(vi)** support the use of the Poisson approximating distribution in part **(iii)**. **[3]**

3 The number of minutes, X, for which a particular model of laptop computer will run on battery power is Normally distributed with mean 115.3 and standard deviation 21.9.

(i) (*A*) Find $P(X < 120)$. **[3]**

(*B*) Find $P(100 < X < 110)$. **[3]**

(*C*) Find the value of k for which $P(X > k) = 0.9$. **[3]**

The number of minutes, Y, for which a different model of laptop computer will run on battery power is known to be Normally distributed with mean μ and standard deviation σ.

(ii) Given that $P(Y < 180) = 0.7$ and $P(Y < 140) = 0.15$, find the values of μ and σ. **[4]**

(iii) Find values of a and b for which $P(a < Y < b) = 0.95$. **[4]**

4 A gardening research organisation is running a trial to examine the growth and the size of flowers of various plants.

(i) In the trial, seeds of three types of plant are sown. The growth of each plant is classified as good, average or poor. The results are shown in the table.

		Growth			Row totals
		Good	Average	Poor	
Type of plant	Coriander	12	28	15	55
	Aster	7	18	23	48
	Fennel	14	22	11	47
Column totals		33	68	49	150

Carry out a test at the 5% significance level to examine whether there is any association between growth and type of plant. State carefully your null and alternative hypotheses. Include a table of the contributions of each cell to the test statistic. **[12]**

(ii) It is known that the diameter of marigold flowers is Normally distributed with mean 47 mm and standard deviation 8.5 mm. A certain fertiliser is expected to cause flowers to have a larger mean diameter, but without affecting the standard deviation. A large number of marigolds are grown using this fertiliser. The diameters of a random sample of 50 of the flowers are measured and the mean diameter is found to be 49.2 mm. Carry out a hypothesis test at the 1% significance level to check whether flowers grown with this fertiliser appear to be larger on average. Use hypotheses $H_0 : \mu = 47$, $H_1 : \mu > 47$, where μ mm represents the mean diameter of all marigold flowers grown with this fertiliser. **[5]**

4767 Statistics 2

Question 1

(i)

x	18	43	52	94	98	206	784	1530
y	1.15	0.97	1.26	1.35	1.28	1.42	1.32	1.64
Rank x	1	2	3	4	5	6	7	8
Rank y	2	1	3	6	4	7	5	8
d	-1	1	0	-2	1	-1	2	0
d^2	1	1	0	4	1	1	4	0

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times 12}{8\times 63}$$

= 0.857 (to 3 s.f.) [*allow* 0.86 to 2 s.f.]

M1 for attempt at ranking (allow all ranks reversed)

M1 for d^2

A1 for $\Sigma d^2 = 12$
M1 for method for r_s

A1 f.t. for $|r_s| < 1$
NB No ranking scores zero

5

(ii)

H_0: no association between X and Y in the population

H_1: some association between X and Y in the population

Two tail test critical value at 5% level is 0.7381

Since 0.857> 0. 7381, there is sufficient evidence to reject H_0,
i.e. conclude that the evidence suggests that there is association between population size X and average walking speed Y.

B1 for H_0

B1 for H_1

B1 for population SOI

NB H_0 H_1 <u>not</u> ito ρ

B1 for ± 0.7381

M1 for sensible comparison with c.v., provided $|r_s| < 1$
A1 for conclusion in words f.t. their r_s and sensible cv

6

(iii)

$\bar{t} = 45$, $\bar{w} = 2.2367$

$$b = \frac{Stw}{Stt} = \frac{584.6 - 270\times 13.42/6}{13900 - 270^2/6} = \frac{-19.3}{1750} = -0.011$$

OR $$b = \frac{584.6/6 - 45\times 2.2367}{13900/6 - 45^2} = \frac{-3.218}{291.6667} = -0.011$$

hence least squares regression line is:

$$w - \bar{w} = b(t - \bar{t})$$
$$\Rightarrow w - 2.2367 = -0.011(t - 45)$$
$$\Rightarrow w = -0.011t + 2.73$$

B1 for $\bar{t}$ and $\bar{w}$ used (SOI)

M1 for attempt at gradient (b)

A1 CAO for -0.011

M1 for equation of line
A1 FT for complete equation

5

(iv)	(*A*) For $t = 80$, predicted speed $= -0.011 \times 80 + 2.73 = 1.85$ (*B*) The relationship relates to adults, but a ten year old will not be fully grown so may walk more slowly. NB Allow E1 for comment about extrapolation not in context	M1 A1 FT provided b < 0 E1 extrapolation o.e. E1 sensible contextual comment	**4**
		TOTAL	**20**

Question 2

(i)	Binomial(5000,0.0001)	B1 for binomial B1 dep, for parameters	**2**
(ii)	n is large and p is small $\lambda = 5000 \times 0.0001 = 0.5$	B1, B1 (Allow appropriate numerical ranges) B1	**3**
(iii)	$P(X \geq 1) = 1 - e^{-} \frac{0.5^0}{0!} = 1 - 0.6065 = 0.3935$ or from tables $= 1 - 0.6065 = 0.3935$	M1 for correct calculation or correct use of tables A1	**2**
(iv)	P(9 of 20 contain at least one) $= \binom{20}{9} \times 0.3935^9 \times 0.6065^{11}$ $= 0.1552$	M1 for coefficient M1 for $p^9 \times (1-p)^{11}$, p from part (iii) A1	**3**
(v)	Expected number $= 20 \times 0.3935 = 7.87$	M1 A1 FT	**2**
(vi)	Mean $= \frac{\Sigma xf}{n} = \frac{7+4}{20} = \frac{11}{20} = 0.55$ Variance $= \frac{1}{n-1}\left(\Sigma fx^2 - n\bar{x}^2\right)$ $= \frac{1}{19}\left(15 - 20 \times 0.55^2\right) = 0.471$	B1 for mean M1 for calculation A1 CAO	**3**
(vii)	Yes, since the mean is close to the variance, and also as the expected frequency for 'at least one', i.e. 7.87, is close to the observed frequency of 9.	B1 E1 for sensible comparison B1 for observed frequency $= 7 + 2 = 9$	**3**
		TOTAL	**18**

Question 3

(i)	(*A*) $P(X < 120) = P\left(Z < \frac{120-115.3}{21.9}\right)$ $= P(Z < 0.2146)$ $= \Phi(0.2146) = 0.5849$	M1 for standardizing A1 for $z = 0.2146$ A1 CAO (min 3 sf, to include use of difference column)	**3**
	(*B*) $P(100 < X < 110) =$ $P\left(\frac{100-115.3}{21.9} < Z < \frac{110-115.3}{21.9}\right)$ $= P(-0.6986 < Z < -0.2420)$ $= \Phi(0.6986) - \Phi(0.2420)$ $= 0.7577 - 0.5956$ $= 0.1621$	M1 for standardizing both 100 & 110 M1 for correct structure in calc^{n} A1 CAO	**3**
	(*C*) From tables $\Phi^{-1}(0.1) = -1.282$ $\frac{k-115.3}{21.9} = -1.282$ $k = 115.3 - 1.282 \times 21.9 = 87.22$	B1 for ±1.282 seen M1 for equation in k and negative z-value A1 CAO	**3**
(ii)	From tables, $\Phi^{-1}(0.70) = 0.5244$, $\Phi^{-1}(0.15) = -1.036$ $180 = \mu + 0.5244\,\sigma$ $140 = \mu - 1.036\,\sigma$ $40 = 1.5604\,\sigma$ $\sigma = 25.63$, $\mu = 166.55$	B1 for 0.5244 or ±1.036 seen M1 for at least one equation in μ and σ and Φ^{-1} value M1 dep for attempt to solve two equations A1 CAO for both	**4**
(iii)	$\Phi^{-1}(0.975) = 1.96$ $a = 166.55 - 1.96 \times 25.63 = 116.3$ $b = 166.55 + 1.96 \times 25.63 = 216.8$	B1 for ±1.96 seen M1 for either equation A1 A1 [Allow other correct intervals]	**4**
		TOTAL	**17**

Question 4

(i)

H_0: no association between growth and type of plant;
H_1: some association between growth and type of plant;

EXPECTED	Good	Average	Poor
Coriander	12.10	24.93	17.97
Aster	10.56	21.76	15.68
Fennel	10.34	21.31	15.35

CONTRIBUTION	Good	Average	Poor
Coriander	0.0008	0.3772	0.4899
Aster	1.2002	0.6497	3.4172
Fennel	1.2955	0.0226	1.2344

$X^2 = 8.69$

Refer to χ^2_4

Critical value at 5% level = 9.488

Result is not significant
There is not enough evidence to suggest that there is some association between reported growth and type of plant;
NB if H_0 H_1 reversed, or 'correlation' mentioned, do not award first B1or final A1

B1 (in context)

M1 A2 for expected values (to 2 dp)
(allow A1 for at least one row or column correct)

M1 for valid attempt at $(O-E)^2/E$
A1 for all correct
NB These M1A1 marks cannot be implied by a correct final value of X^2

M1 for summation
A1 for X^2 CAO

B1 for 4 d.o.f.

B1 CAO for cv

M1
A1

12

(ii)

Test statistic = $\dfrac{49.2-47}{8.5/\sqrt{50}} = \dfrac{2.2}{1.202} = 1.830$

1% level 1 tailed critical value of z = 2.326

1.830 < 2.326 so not significant.
There is not sufficient evidence to reject H_0

There is insufficient evidence to conclude that the flowers are larger.

M1 correct denominator
A1

B1 for 2.326
M1 (dep on first M1) for sensible comparison leading to a conclusion

A1 for fully correct conclusion in words in context

5

TOTAL **17**

4767 Statistics 2

General Comments

For the majority of candidates this proved to be a straightforward paper, with many high marks achieved. Most candidates demonstrated a good ability to carry out statistical tests and interpret results using appropriate language. It is pleasing to see candidates providing conclusions to their hypothesis tests which are not 'too assertive'; this is a requirement in Statistics 3 but, at the moment, some flexibility is allowed in Statistics 2. On the whole, candidates scored well on all questions, but question 2 provided the toughest challenge.

Comments on Individual Questions

Section A

1) (i) Candidates were required to find the value of Spearman's rank correlation coefficient from raw data. This produced full marks for most candidates. In addition to numerical mistakes, common errors included incorrect application of the formula - omitting 6 from $6\times\Sigma d^2$ and failing to use '1 - …' were often seen. Very few candidates failed to attempt to rank the data.

(ii) Most candidates scored well. On the whole, hypotheses were stated correctly, using the appropriate form of null hypothesis – H_0: No association. Most candidates obtained the correct critical value, sensibly compared their test statistic from part (i) and made an appropriate conclusion. In keeping with previous sessions, the most common reason for losing marks involved failing to carefully specify the hypotheses, in context, to make it clear that the test was for association between city population size and average walking speed of pedestrians in the population.

(iii) This part was well answered, with many candidates awarded full marks. In some cases, marks were lost through inaccurate working – e.g. giving the value of the gradient of the regression line correct only to one significant figure. Several candidates used x and y instead of w and t. Those candidates who obtained a positive value for the gradient of the regression line (which was clearly shown on the question paper as having a negative gradient) were more heavily penalised than those making minor errors in their calculation of the gradient.

(iv) (A) Well answered. Most candidates gained both marks and were able to make a sensible comment in part (B).

(iv) (B) Most candidates realised that using the regression line to estimate the maximum walking speed of a 10-year-old male constituted 'extrapolation' – some went on to provide comments that, in this case, it was not sensible due to physical development issues. Those candidates who provided a statistically based comment together with a pertinent contextual comment generally picked up both available marks.

2 (i) Well answered. Most candidates gained both marks. Some candidates jumped ahead, stating that the distribution was Poisson, making it difficult to 'explain why X may be approximated by a Poisson distribution in part (ii).

(Ii) Well answered. Most candidates were awarded all three marks. Some candidates covered all bases, providing general comments to justify use of a Poisson distribution in its own right in addition to those supporting the Poisson approximation to the Binomial distribution. Numerical mistakes were rare.

(iii) Fully correct answers were plentiful. Few candidates found P(X = 0) rather than 1 - P(X = 0). Some used 1 - P(X = 1), which scored no marks.

(iv) Many candidates scored full marks. Use of X ~ Po(10) was seen regularly, leading to 0 marks for this part of the question.

(v) Well answered. Many candidates felt the need to provide an integer answer; this was condoned provided that 7.87 was seen.

(vi) Most candidates correctly obtained 0.55 for the mean of the data provided. Attempts to calculate the variance were poor with many failing to use the (n – 1) divisor as required.

(vii) Many candidates lost marks through failing to provide comments relating to the answer to part (v). Candidates were required to compare the expected number of samples containing at least one four-leaf clover with the observed number and to provide numerical values to show that they were comparing appropriate values. Very few good answers were seen. Many scored a mark for a sensible comment relating to their values for the mean and variance found in part (vi), although some compared mean with standard deviation.

3 (i) (A) Most candidates obtained full marks. Very few lost a mark by failing to work with the sufficient accuracy (i.e. making use of the 'difference' column in the Normal tables), even fewer failed to standardise correctly, commonly dividing by $\sqrt{\sigma}$ or σ^2. However, many attempts at continuity corrections were seen.

(i) (B) Many fully correct answers seen. Most managed to correctly standardise the ends of the given inequality, but many candidates made mistakes with the structure of the required probability calculation. Those applying continuity corrections lost at least the mark for accuracy.

(i) (C) Well answered by many. Many candidates used a positive z value leading to a value of k greater than the mean, leading to a maximum of 1 mark out of 3. Those failing to use inverse Normal tables (i.e. those using probabilities in place of z values) were awarded no marks.

(ii) Many candidates scored full marks. However, use of +1.036 instead of -1.036 was common and lead to a negative value for σ; despite this, most candidates did not spot their error. Attempts to solve simultaneous equations were, on the whole, good; however, those failing to use inverse Normal tables scored no marks.

(iii) This part caused problems for many, not least identifying suitable z values. Many candidates struggled to use their z value(s) in appropriate equations. However, many correct answers were seen (including non-symmetrical intervals).

4 (i) Most candidates scored a mark for providing correct hypotheses, although some failed to write the hypotheses in context.
Despite being asked specifically, many candidates failed to 'include a table of the contributions of each cell to the test statistic', making it difficult to award marks for accurate working as often only the final X^2 value was given. Most candidates identified the correct number of degrees of freedom and the corresponding critical value, then went on to make an appropriate conclusion to the test. As mentioned in the general comments, it is encouraging to see phrases such as 'the evidence suggests that', rather than 'this proves that', appearing with increasing regularity.

(ii) Well answered. In previous years, many candidates have failed to use the sampling distribution of means in their calculation of the test statistic when tackling questions such as this. This year, the vast majority of candidates scored well in this part. Some lost a mark for stating an incorrect critical value. Others lost marks for inappropriate comparisons (typically, comparing a z value with a probability). On the whole, the wording used in conclusions to hypothesis tests has vastly improved compared to previous years, although there are still some candidates who do not use any context in their conclusions and are penalised.